

EDGE CONFLICTS DO NOT DETERMINE GEODESICS IN THE ASSOCIAHEDRON

SEAN CLEARY AND ROLAND MAIO

ABSTRACT. There are no known efficient algorithms to calculate distance in the one-skeleta of associahedra, a problem which is equivalent to finding rotation distance between rooted binary trees. One approximate measure of distance in associahedra is the extent to which the edges in the attached triangulations are incompatible, and a natural way of trying to find shortest paths is to maximize locally the number of compatible edges between triangulations. Such steps minimize the number of conflicting edges between these triangulations. We describe examples which show that the number of conflicts does not always decrease along geodesics. Thus, a greedy algorithm which always chooses a transformation which reduces conflicts will not produce a geodesic in all cases.

1. INTRODUCTION

Rooted binary trees arise across a range of areas, from phylogenetic trees representing genetic relationships to efficient organizational structures in large datasets. When considering two rooted binary trees, there are a wide range of possible measures of distance between them, depending upon the situation and how much structure we attach to the trees. In the setting of rooted binary trees with a natural right-to-left order, such as those corresponding to binary search trees, a widely-considered distance is rotation distance. The *rotation distance* $d(S, T)$ between two trees S and T is the minimum number of rotations needed to transform the tree S to the tree T . There is a natural bijection between rooted binary trees with n leaves and triangulations of a marked regular polygon with $n + 2$ sides. The operation that corresponds to rotation at a node for binary trees corresponds to an edge-flip between

Key words and phrases. random binary trees.

Partial funding provided by NSF #1417820. This work was partially supported by a grant from the Simons Foundation (#234548 to Sean Cleary).

triangulations. Thus, rotation distance between two trees is exactly equivalent to a corresponding edge-flip distance between triangulations.

The associahedron of size n is a combinatorial object capturing many aspects of triangulations, trees, or possible associations of expressions. Here, by the associahedron of size n we mean a graph whose vertices are triangulations of a marked regular $n + 2$ -gon (equivalently, rooted trees with $n + 1$ leaves) and where two vertices S and T are connected if there is a single edge flip which transforms the triangulation S to T (equivalently, if the associated rooted trees differ by a single rotation at a node.) This is the one-dimensional skeleton of potentially higher-dimensional descriptions also known as associahedra. Here, we work entirely in the one skeleton and neglect the additional higher dimensional structure of associahedra realized as convex polytopes.

There are no known polynomial-time algorithms to find shortest paths in one-skeleta of associahedra or to find distances between vertices in those one skeleta. There are a number of approximation algorithms [1, 4] for rotation distance. Two edges are said to be in conflict if it is not possible for them to be part of the same triangulation, due to the two edges crossing. This is the ordered version of having two edges be in conflict in trees without an order on leaves, such as those arising in phylogeny. Edges which are not in conflict are said to be compatible, and two trees which have a large number of compatible edges are considered reasonably close. See Semple and Steel [7] for discussion of combinatorial and algorithmic questions about counting the number of compatible edges (or the number of edge conflicts) between phylogenetic trees, including estimates of distance from increasing the number of compatible edges, which is equivalent to reducing the number of conflicting edges.

Here, we investigate the potential connection between edge conflicts and finding geodesics in the setting of ordered trees, or equivalently of triangulations of polygons. Two triangulations which have many edges which are in conflict are generally far from one another, and triangulations which have few conflicts are generally quite close together. The number of expected conflicts between triangulations selected uniformly at random has been analyzed experimentally by Chu and Cleary [2]. Here, we describe examples where there is no geodesic along which the total number of conflicts between the triangulations uniformly decreases or even remains the same. This rules out the potential success

of a greedy algorithm to reduce conflicts to find shortest paths or distances in associahedra.

2. BACKGROUND AND DEFINITIONS

We consider the marked regular $n + 2$ -gon P with edges labelled as R for root and then consecutively from 0 to n . A *triangulation* T of P is a collection of $n - 1$ non-crossing edges from vertices of P which separate P into n triangles. An *edge flip on T* is the process of taking two triangles which share an interior edge, thus forming a quadrilateral Q , and replacing the interior diagonal of Q which lies in T with the other diagonal of Q to form a new triangulation T' . For each k , the relevant *associahedron* is a graph whose vertices are triangulations of the regular $k + 2$ polygon and whose edges connect vertices which differ by a single edge flip. The *edge flip distance* from a triangulation S to a triangulation T of the same size is the minimal length path in the associahedron. Alternatively, we can regard the vertices of the k associahedron as rooted binary trees of size k interior nodes with $k + 1$ exterior nodes (known as leaf nodes) numbered from 0 to k , with the edges connecting vertices whose trees differ by a single rotation (left or right) at an interior node. To any triangulation, there is a natural dual construction giving rise to the tree description. Edge flip distance between triangulations is known as *rotation distance* between binary trees.

Rotation distance was described by Culik and Wood [5] whose arguments showed that the one-skeleta of associahedra are connected and gave an upper bound of $2n - 2$ as the distance between any two vertices of the associahedron of size n . Remarkable work of Sleator, Tarjan and Thurston [8] showed that the upper bound for distance between vertices for $n \geq 11$ is $2n - 6$, and furthermore that that upper bound is realized for all n larger than some very large N . Recent work of Pournin [6] showed that in fact the upper bound is realized for all $n \geq 11$. There are no known polynomial-type algorithms to compute rotation distance, although rotation distance has been shown to be fixed parameter tractable by Cleary and St. John [3] and there are a number of approximation algorithms [1, 4].

We describe trees via the binary sequence obtained by a pre-order traversal all nodes of the tree, recording a 1 for each internal node and 0 for each leaf node, giving a binary sequence of n 1's and $n + 1$ 0's

via this encoding. So for example the balanced tree with four leaves is encoded as 1100100. For the figures below, we draw triangulations as collections of chords in the hyperbolic disk with the boundary as an $n + 2$ -gon, with the intervals numbered counterclockwise from 0 to n and the root interval unlabelled at the top. The counterclockwise end of the interval labeled i is referred to as vertex i when needed.

Two edges s and t are said to be *conflicting* if it is not possible for them to be present in the same triangulation; equivalently, if the edges cross as chords connecting vertices in the relevant polygon. Two edges which are not in conflict are said to be *compatible*, a term used widely in phylogenetic settings. For example, a chord from vertex 5 to 8 is in conflict with a chord from vertex 3 to 6, which can be seen as one of the intersections of the red chords and blue chords in Figure 2 if we number the vertices as the counterclockwise ends of the numbered edges. The number of conflicts for a pair of triangulations S and T is the sum of the conflicts between the pairs of edges. We note that the total number of conflicts between S and T is an estimate of the distance between S and T only weakly. In particular, the number of conflicts between S and T is a semi-metric in that it is symmetric and positive definite and satisfies that if the total number of conflicts is zero, the triangulations must coincide. However, it is not a metric in that it does not satisfy the triangle inequality.

3. CONFLICTS ALONG GEODESICS

There are a range of conflict-based greedy algorithms to reduce conflicts to try to find a geodesic from a given triangulation S to a given target triangulation T . Generally, they proceed in a manner under the following approach:

- (1) Set $S = S_0$ to begin
- (2) If S_i is T , then we are done and the distance is no more than i .
- (3) If S_i is not the triangulation T , then we calculate the conflicts between all of the neighbors of S_i and the target T . Among those neighbors, we let S_{i+1} be a tree with minimal conflicts, and then we repeat.

The simplest version is merely to always take the neighbor with minimal conflicts which is lexicographically least. There are other more sophisticated approaches for choosing among ties between neighbors

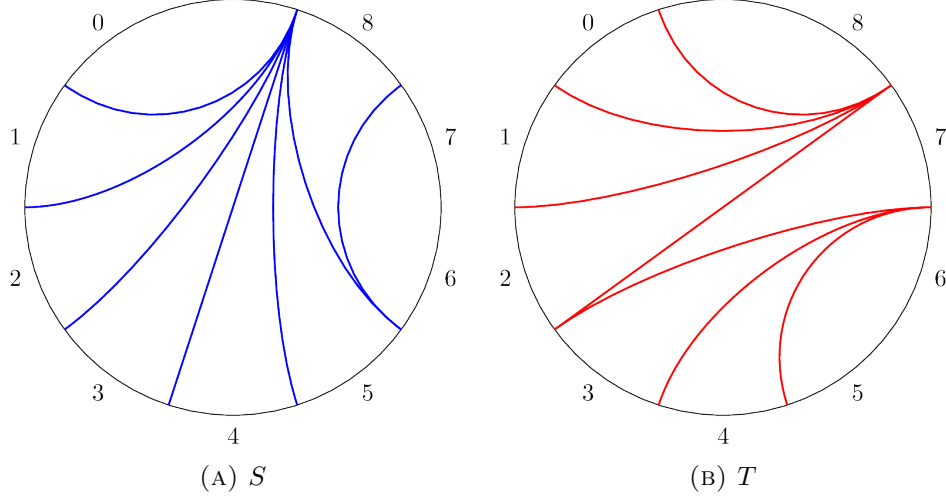


FIGURE 1. The pair of triangulations S and T used to prove Proposition 1.

using some kind of greedy approach but we will see below that any of them are guaranteed to give an overestimate in some cases.

We note that this algorithm will produce a path from S to T since unless the trees coincide, there is always a neighbor with fewer conflicts with the target tree. Thus the algorithm will always give a path from S to T , and thus gives an upper bound for the edge-flip distance. However, this path is not necessarily a geodesic path so the distance from S to T may be less than various greedy conflict-based algorithms may find. There are multiple ways in which such a greedy algorithm to reduce conflicts can fail to find a geodesic. It may be that there are several choices among the neighbors of S_i with the same number of conflicts, and at least one of them does not begin a geodesic path from S_i to T . Or it may be that none of the neighbors with minimal conflicts begins a geodesic path. Broadening the notion to allow choices where the number of conflicts is not necessarily minimal but merely equal or smaller than the current number of conflicts gives many more possibilities. But even algorithms which consider reducing conflicts (not necessarily minimally) or keeping conflicts constant will not always find a geodesic path because of examples of the following type:

Theorem 1. *There are examples of triangulation pairs (S, T) where every geodesic γ from S to T with $\gamma = \{S = S_0, S_1, S_2, \dots, S_k = T\}$ has the property that S_1 has more conflicts with T than S has with T .*

Proof. We consider a pair of triangulations S and T , each with 7 chords dividing the 10-gon into 8 triangles. Their dual trees correspond to trees with 8 internal nodes and 9 leaves. The partition S has encoding 10101010101011000 for its dual tree and T has encoding 11010101101010000.

The triangulations S and T are shown in Figure 1. The distance in the associahedron is 8, obtained by a breadth-first enumeration of neighborhoods of increasing size of S . The number of conflicts between S and T is 27 total conflicts, as shown in Figure 2 where the trees are superimposed.. There are 7 neighbors to S , each corresponding to flipping one of the seven edges. The number of conflicts of those neighbors of S with T are 27, 26, 26, 23, 22, 22 and 28. The only neighbor of S which is closer to T than S is the last one, which has more conflicts than the original S . Thus any geodesic from S to T will necessarily have the conflicts rise from 27 to 28. Note that we can add identical additional triangles to each of the triangulations in the pair to create examples of any size at least 8. \square

Thus, any algorithm which attempts to find geodesics by either minimizing conflicts, at least reducing conflicts, or keeping conflicts at least non-increasing will not find any geodesic from S to T .

We note a few properties of this particular example. Though it is quite common to have multitudes of geodesics between a pair of trees, there is in this case a unique geodesic γ from S to T . This geodesic proceeds through triangulations listed in Table 1 which are pictured in Figure 3.

This geodesic γ begins with a first move which actually has the maximum number of conflicts among all neighbors of S . The greedy algorithm, taking the lexicographically first minimal conflict neighbor in each case, gives a path of length 9, whereas the minimal length path is of length 8, thus giving an over-estimate by 1.

This particular example is not symmetric. We note that if we proceed instead from T toward S , the neighbors of T have conflicts 21, 23, 24, 29, 25, 25, and 26 with S , with the only neighbor of T which is closer to S having 21 conflicts, the smallest of all choices with respect to conflicts. And, in fact, the greedy algorithm which proceeds from the T toward S in steps by always taking the lexicographically least neighbor with the fewest conflicts with S does find a geodesic path of length 8.

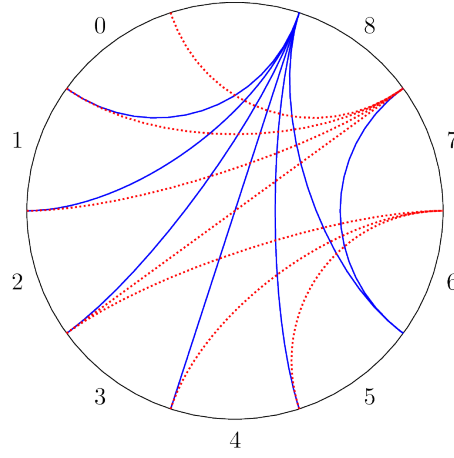


FIGURE 2. The triangulations S and T superimposed, with S in blue and T in dotted red. There are 27 conflicts in the pair (S, T) which can be seen as the 27 intersections between the triangulations

| Tree | Conflicts with T |
|---------------------------|--------------------|
| $S = S_0$ | 27 |
| $S_1 = 10101010101010100$ | 28 |
| $S_2 = 10101010101100100$ | 21 |
| $S_3 = 10101010110100100$ | 15 |
| $S_4 = 10101011010100100$ | 10 |
| $S_5 = 10101011101010000$ | 6 |
| $S_6 = 10101101101010000$ | 3 |
| $S_7 = 10110101101010000$ | 1 |
| T | 0 |

TABLE 1. The triangulations in the geodesic γ from S to T and the conflicts with the target tree T .

However, by merely concatenating the two tree pairs (S, T) and (T, S) along any of the peripheral edges to obtain a pair of triangulations (U, V) of the 19-gon which is symmetric with respect to the fact that for any geodesic γ from U to V or from V to U , there must be at least one point along γ where the total number of conflicts must increase in either direction.

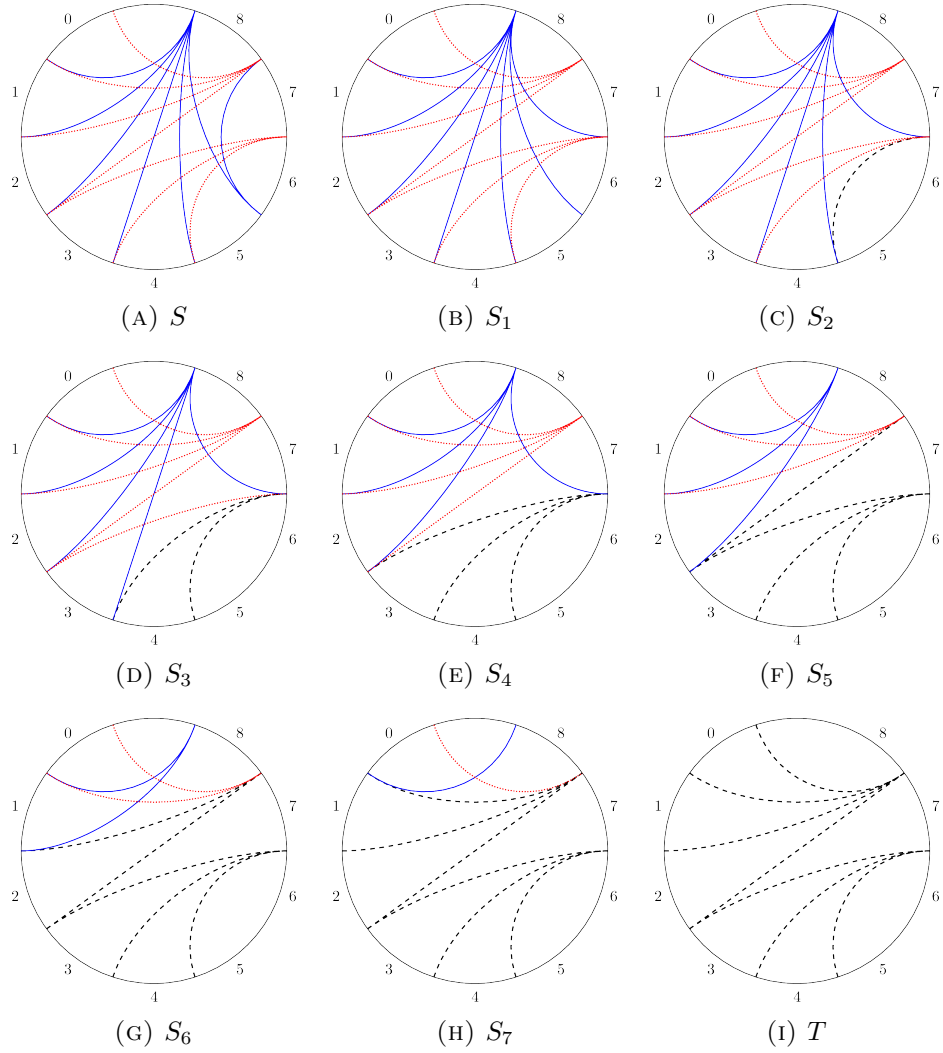


FIGURE 3. The triangulations in the geodesic γ from S to T superimposed on the target T . The triangulation S is in solid blue and T in dotted red, and common edges are shown as dashed black.

The example pairs S and T given above are one of 28 equivalence classes (up to rotational and reflectional symmetry of the triangulations) of examples of size 8 where the initial geodesic step must increase the number of conflicts.

We do note that this kind of behavior where every geodesic begins with an increase in conflicts are somewhat rare, with these 28 equivalence classes of conflict-increasing examples of size 8 occurring from among the 117,260 equivalence classes of edge-flip distance problems of size 8. For size 9, the fraction is also small- there are 632 equivalence classes where all geodesics have an increase in conflicts, out of 1,108,536 classes of problems.

Though the examples above show that conflict-reducing algorithms do not always give the correct distances, they often do give correct distances and when they do, typically the gap between the the simplest conflict-based estimate and the true geodesic distance is generally not large.

In the case of the 117,260 size 8 edge-flip distance problems of which (S, T) is one instance, the lexicographically-least minimal conflict greedy algorithm is correct in 111,061 cases, with 5771 cases overestimating the distance by 1, 423 cases of an overestimate by 2, and the furthest it is ever off by is 3 which happens in 5 cases. That results in about a 6% overestimate of distance on average across all problems of size 8.

For the 1,108,536 equivalence classes of distance problems of size 9, despite the 632 cases where the all geodesics begin with an increases in conflicts, the naive greedy algorithm gives the geodesic distance correctly in more than 1 million cases. In the cases where it is incorrect, in 4 cases the overestimate of the distance is 4, in 444 cases the overestimate is 3, in 7047 cases it is off by 2, and in 80,710 cases the algorithm overestimates by 1. On average, this greedy algorithm overstates the distance by about .0867378. Similarly, in the case of size 10 for triangulations of the 11-gon, the vast majority of distances (89%) are correctly determined with an average of 0.11888 overestimate of the distance.

REFERENCES

- [1] Jean-Luc Baril and Jean-Marcel Pallo. Efficient lower and upper bounds of the diagonal-flip distance between triangulations. *Information Processing Letters*, 100(4):131–136, 2006.
- [2] Timothy Chu and Sean Cleary. Expected conflicts in pairs of rooted binary trees. *Involve*, 6(3):323–332, 2013.
- [3] Sean Cleary and Katherine St. John. Rotation distance is fixed-parameter tractable. *Inform. Process. Lett.*, 109(16):918–922, 2009.
- [4] Sean Cleary and Katherine St. John. A linear-time approximation for rotation distance. *J. Graph Algorithms Appl.*, 14(2):385–390, 2010.

- [5] K. Culik and D. Wood. A note on some tree similarity measures. *Information Processing Letters*, 15(1):39–42, 1982.
- [6] Lionel Pournin. The diameter of associahedra. *Adv. Math.*, 259:13–42, 2014.
- [7] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.
- [8] Daniel D. Sleator, Robert E. Tarjan, and William P. Thurston. Rotation distance, triangulations, and hyperbolic geometry. *J. Amer. Math. Soc.*, 1(3):647–681, 1988.

DEPARTMENT OF MATHEMATICS, THE CITY COLLEGE OF NEW YORK AND
THE CUNY GRADUATE CENTER, CITY UNIVERSITY OF NEW YORK, NEW YORK,
NY 10031

E-mail address: cleary@ccny.cuny.edu

URL: <http://www.sci.ccny.cuny.edu/~cleary>

DEPARTMENT OF COMPUTER SCIENCE, THE CITY COLLEGE OF NEW YORK,
CITY UNIVERSITY OF NEW YORK, NEW YORK, NY 10031

E-mail address: rolandmaio38@gmail.com